Compressed Domain Motion Analysis for Video Semantic Events Detection

Kun Tao
Graduate School of
Chinese Academy of Sciences
Beijing, China
ktao@ict.ac.cn

Shouxun Lin, Yongdong Zhang
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China
{sxlin, zhyd}@ict.ac.cn

***Abstract*—In this paper, a novel approach is proposed to estimate camera motion and segment moving objects from compressed video streams, aiming to detect semantic events in video clips. Simultaneously using the motion vectors and DC components of MPEG macroblocks (MB), the camera motion type and motion parameters of each frame are estimated with simplified models. Then the segmentation of moving objects is done at macroblock level. Exploring the variation of motion information in consecutive frame sequence, a composite feature can be formed to detect semantic events. The experiment results on TRECVID video corpus show that our approach is very effective and efficient.**

Keywords-compressed domain; motion analysis; object segmentation; semantic event; mpeg

I. INTRODUCTION

For the developing applications of digital videos, efficient methods to analysis video contents will be increasingly important. The semantic contents of videos include objects, scenes and events etc., which can be detected from key-frame images, audio and other information of different modalities. The motion information, including global camera motion and moving objects in video scenes, is very valuable for detecting semantic events such as Walking, Violence or People-Marching. Most existing works on motion analysis are based on visual features of frame images [1,2], that means a time-consuming uncompressing process is needed. For the applications on large-scale video corpus, a faster compressed domain method is necessary.

The most important basis of semantic events detection is the extraction of motion information, including camera motion estimation and moving objects segmentation. In last decades, a few compressed domain methods have been proposed, which use the motion vectors and/or DCT coefficients directly to avoid the IDCT process and motion compensation. Although these methods perform well on some special cases, they all have to face some difficulties in general applications: First, the DCT coefficients of inter-coded MBs are not calculated from the real pixel data, but from the residual errors of pixel data between current MB and its reference region. Second, there are many intra-coded MBs in video streams which can not provide motion information. Especially, I-frames are totally intra-coded. The last and most important problem is that motion vectors are highly noisy. Motion vectors of MPEG videos are calculated by a fast macroblock matching algorithm in encoder, which may not accord with the real motion of objects, especially in uniform (non-textured) regions. The solutions of these problems can strongly influence the performance of a compressed domain method.

Some existing methods can work only with the hypothesis of Still or Pan camera motion [3-5], and some try to estimate a 6 parameters affine model or 8 parameters projective model [6]. A typical parameter method named iterative rejection scheme was proposed in [7] and improved by [8] and [9]. This method iteratively estimates the bilinear global-motion model by least-square estimation and rejects outlier MBs whose motion vectors result in larger than average estimation errors. The rejected MBs are regarded as local (object) motion, and others which accord with the global motion belong to background. Temporal consistency constraints can be applied to discriminate actual moving objects from noise MBs [3,4,8-10]. Then some descriptors and the trajectories of segmented moving objects are counted for applications such as retrieval [1,7,8] or semantic events detection[2].

Our approach tries to extract the motion information with simplified models. In [11] an effective method is proposed to discriminate different kinds of camera motion, which is inherited and developed by us for motion segmentation. Besides the reduction of computation time, our method has another two advantages: A) As most existing methods discard the information of I-frames [3,4,7,10], we use a special preprocessing step to reevaluate the potential real motion of intra-coded MBs in I-frames; B) We use the approximations of real DC components for inter-coded MB[12], which makes our method more reliable than those who use residual error DC components directly. Different from temporal consistency constraints, another trait of our method is that we filter out noise MB based on motion vectors and DC components in every isolated frame. Some invariant moments [13] can be calculated from the mask image of segmentation, and be formed into the feature of a frame sequence for semantic events classification. The proposed approach has been used in our system for 2007 TRECVID evaluation [14], and been proved to be valuable by experiments. The details of the camera motion estimation and moving object segmentation are described in Section 2. Section 3 is about the semantic events detection. Experiment results and conclusions are in Section 4 and 5 respectively.

II. CAMERA MOTION ESTIMATION & MOVING OBJECTS SEGMENTATION

A. The Preprocessing Step

The motion vectors used in our approach are extracted from compressed MPEG streams. First of all, the extracted motion vectors should be scaled to make them independent of frame type. All motion vectors are divided by the difference between current frame number and the reference frame number (in the display order). Such step makes them close to the magnitude of the real motion between adjacent frames. Then the signs of backward motion vectors are revised. In the case of bidirectionally predicted MBs, the forward motion vectors and revised backward motion vectors are averaged.

The reevaluation of I-frames' motion information relies on their adjacent B-frames. Usually a certain I-frame acts as the reference frame of at least one adjacent B-frame. If a MB in I-frame overlaps the best matching region of an inter-coded MB in adjacent B-frames, the motion vector of the latter will be regarded as an approximation to the motion of the former. In case a MB in I-frame corresponds to several MBs in B-frames, only the one with the highest overlap rate will be taken into account. If such corresponding MBs can not be found, the MB in I-frame will keep being regarded as intra-coded.

Before estimating the global camera motion, the noise MBs with abnormal motion vectors should be eliminated. Using DC components extracted by the method of [12], the MBs which meet the following three criterions are marked as noise: a) the motion vector of current MB obviously differed from the motion vectors of neighbor MBs, b) the DC values of four 8*8 blocks inside current MB are nearly the same, c) more than half of 8 neighbor MBs have DC values close to the DC values of current MB. Both the noise MBs and the intra-coded MBs will be treated distinguishingly in following steps.

B. Global Camera Motion Estimation

Instead of estimating affine model or projective model, we judge different camera motion types with several simplified discrimination models. Then for different types of camera motion, the corresponding model parameters will be calculated.

The first two steps are to discriminate the modes Still and Pan/Tilt. All motion vectors are transformed to polar coordinate except for noise or intra-coded MBs. The number of all MBs, noise MBs and intra-coded MBs in current frame can be defined as n_{all}, n_{noise} and n_{intra}. If the number of MBs with zero polar radiuses is larger than $0.4*n_{all}$, the frame will be regarded as Still. Once current frame does not belong to Still, the following step is Pan/Tilt discrimination. The polar radiuses are rounded to nearest integers, and the polar angles are also denote by integers between $[0,360)$. Then a 2-D histogram is constructed from polar radiuses and polar angles. Selecting a point (ρ,θ) in 2-D histogram as center point, the summated value is counted on the neighborhood window. Sliding the window on whole histogram for the largest summated value, the resulting center $(\hat{\rho},\hat{\theta})$ can indicate the majority motion. If the largest summated value $n_{major} > (n_{all} - n_{noise})/2$, the camera motion is regarded as Pan/Tilt. For a frame whose width and height in pixel are w and h, $\hat{\rho}/(0.5*w)$ will be defined as the global motion magnitude parameter r.

Whenever the discrimination results of Still and Pan/Tilt are negative, the next two steps are discriminations of Zoom and Rotation. A matrix of the same size of frame image is built up first. At each pixel point a continue line is drawn in the direction of the motion vector, except for the points in noise or intra-coded MBs. Every time being passed by a continue line, the corresponding element in matrix increment by one. During a zoom motion of the camera, ideally all motion vectors should point to the focus-of-expansion (FOE). Therefore the elements around the FOE should be very large. By searching for a 5*5 window with largest summated value in accumulation matrix, a potential FOE is decided. If the largest summated value is 10 times larger than the average, the camera motion is regarded as Zoom. Then for each pixel point, d_{FOE} denotes its distance to the FOE. We calculating the value ρ / d_{FOE} at all points whose motion vectors approximately aim at FOE. The average of ρ / d_{FOE} is regarded as the magnitude parameter r. We use $r > 0$ to indicate Zoom In, and $r < 0$ to indicate Zoom Out. Figure 1 can illustrate the motion vectors and matrix distribution of a Zoom frame. The discrimination of Rotation is almost the same except that the continue lines should be drawn in the vertical direction of the motion vectors. $r > 0$ indicates clockwise rotation, and $r < 0$ indicates anticlockwise rotation. Finally, if all four discrimination results are negative, or $n_{noise} + n_{intra} > n_{all}/2$, the frame is labeled as Irregular.

Experiments on 2007 TRECVID development corpus show that 95.8% of more than 6 million frames are regarded as Still, 3.6% are Pan/Tilt, 0.6% are Zoom, Rotation and Irregular. So despite that the discrimination steps of Zoom and Rotation are somewhat time-consuming, they are rarely used.

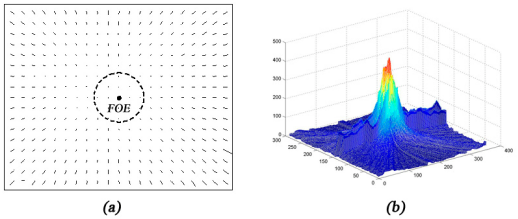

Figure 1. Zoom motion. (a)the motion vectors (b)accumulation matrix

C. Segmentation

Having got the global camera motion type and corresponding motion parameters, the camera motion compensation can be made. For Pan/Tilt, Zoom or Rotation frames, we can estimate the camera motion vectors caused by camera motion at all pixel points except the points in noise or intra-coded MBs. In a Pan/Tilt frame, camera motion vectors are all equal to $(\hat{\rho},\hat{\theta})$. In a Zoom frame, a camera motion vector is in the same direction of the continue line between current pixel point and the FOE. The orientation is decided by the sign of r, and the polar radius is $|r|*d_{FOE}$. The camera motion vectors in Rotation frames can be calculated in similar way. The camera motion vector of the center point is selected to denote the motion of a whole MB. Subtracting the camera motion vector from the MB's motion vector, the result indicates the relative motion of the MB to global background. The MB whose relative motion has a polar radius larger than the empirical threshold 2.0 can be labeled as foreground (moving objects).

Noise MBs and intra-coded MBs will not be compensated. If more than half of their 8 neighbors are normal MBs, they obey the majority of normal neighbors' foreground/background label. Otherwise they will be labeled as background. Then a spatial median filter will be applied at MB magnitude to remove isolated moving MBs. As a result, the final segmentation mask can be gained. The camera motion compensation and segmentation result are illustrated in Figure 2.

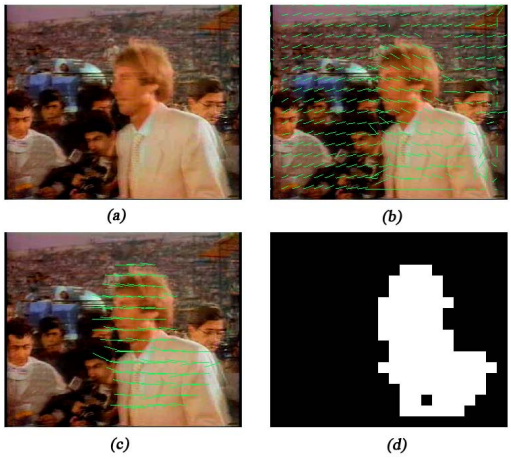

Figure 2. Camera motion compensation and segmentation. (a)original frame image (b)macroblock motion vectors (c)motion vectors after compensation and thresholding (d)segmentation mask

III. Semantic Events Detection

A mask image can reveal the size, shape and location of moving objects. In our approach, the mask image can be characterized as some statistical moments. Following the symbol definition in [13], (x_c, y_c) is defined as the gravity center of foreground regions, m_{00} is the region area, and u_{pq} is a translation invariant moment of order $(p+q)$. Specially, in our approach the pixel value $f(i,j)=1$ in foreground MBs and $f(i,j)=0$ in background. All 7 u_{pq} moments of order 2 and order 3 can be calculated to represent the information about size, orientation and shape of foreground. The other 7 rotation-, translation-, and scale-invariant moments $\varphi_1,\cdots,\varphi_7$ in [13] are also used, which can characterize the inherent shape of foreground regions.

According to the frame image size, m_{00} and (x_c, y_c) can be scaled to $m_{00}/(w*h)$, x_c/w and y_c/h. The 7 u_{pq} moments are also divided by $(w*h)$. Together with $\varphi_1,\cdots,\varphi_7$, all above parameters are formed into a 17-D feature $\vec{V}$ which can characterize the information of mask image, i.e. the information of moving objects.

As semantic events are detected with a shot as the unit, the segmentation results should be farther explored in the shot frame sequence. The semantic events usually occurred only on a portion of the shot, so we detect them in a sliding window. If an event is detected in the window, it is also contained by the whole shot. Since the length of the shots in TRECVID 2007 corpus distributes from 2 seconds to 5 minutes, a window of 2 seconds is used in our approach. The feature sequence of a window is $\{\vec{V}_1,\vec{V}_2,\cdots,\vec{V}_n\}$, whose average is $\hat{\vec{V}}$. The absolute values of neighbor difference $\{\Delta\vec{V}_1,\Delta\vec{V}_2,\cdots,\Delta\vec{V}_{n-1}\}=\{|\vec{V}_1-\vec{V}_2|,|\vec{V}_2-\vec{V}_3|,\cdots,|\vec{V}_{n-1}-\vec{V}_n|\}$ are also calculated, whose average is $\Delta\hat{\vec{V}}$. $\Delta\hat{\vec{V}}$ is more recommendable than variances in this application because it includes the information of temporal variation. Other four statistics will be aloes counted on the window, including p_{null} (the percent of frames which have no moving objects), p_{Still} (the percent of Still frames), $p_{Irregular}$ (the percent of Irregular frames) and $\hat{r}$ (the average of all existing motion magnitudes r). $\hat{\vec{V}}$, $\Delta\hat{\vec{V}}$ and above four statistics form the final 38-D feature. Different for those complex models, our feature can be directly used as the input of a SVM classifier to detect different semantic events. The main modules of our approach are shown in Figure 3.

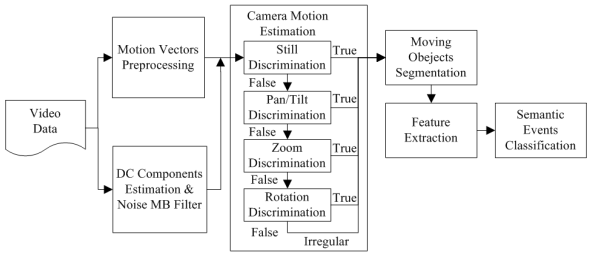

Figure 3. Overview of the proposed approach

IV. EXPERIMENT RESULTS

High-level feature extraction is a primary task of TRECVID evaluation, which aims to detect various high-level semantic concepts. A 39-concepts lexicon related to events, objects, locations, people and programs is employed. Then the concept labels of development corpus are also provided. Our experiments were carried out on the latest 2007 development corpus. More than 18,000 shots were divided into a training subset and a testing subset randomly. A baseline approach was used in contrast, which is based on six kinds of visual features extracted from key-frames. Comparing the results of the proposed approach, baseline and the fusion results of them, we found that our approach is good at detecting concepts about events or behavior such as "Walking_Running" and "People-Marching", and can boost the detection precision for other few concepts such as "Sports" by fusion with the baseline. The precision-recall curves of "Walking_Running" and "Sports" are shown in Figure 4.

The total length of the videos in development corpus is about 70.9 hours, all above steps from motion vector extraction to feature calculation were finished in about 52.4 hours, faster than real-time. Such experiments were performed on a machine with AMD Dual Core 265 Processor. Then for concepts having been proved to benefit from motion information, the prediction results based on motion features will be combined with results based on other modalities in our system. As examples, one of our submitted runs with motion analysis ranks 6th among all 163 runs for the concept "Sports" in official reports, and 35th for "People-Marching".

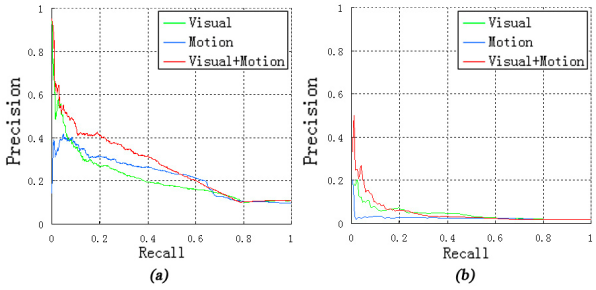

Figure 4. The precision-recall curves of visual baseline, motion and the fusion result for (a)Walking_Running (b)Sports

V. CONCLUSION

In this paper, we proposed a novel approach which can extract motion information from compressed video streams and detect semantic events in video shots. The low time consumption and good effect make our approach favorable to the applications on large-scale video corpus. Experiments on 39 concepts proposed by TRECVID show that our approach is very effective for those concepts about events or behaviors. By fusion with different features, it can also contribute to the detection of some other concepts. Having been proved valuable in our current work, its potential in other applications is also worth to be expected.

ACKNOWLEDGMENT

This work was supported in part by the National Basic Research Program of China (973 Program, 2007CB311100), the National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), National Nature Science Foundation of China (60873165 、 60802028), the Beijing New Star Project on Science & Technology (2007B071), and Co-building Program of Beijing Municipal Education Commission.

REFERENCES

[1] S. Dagtas, W. Al-khatib, A. Ghafoor, and R. L. Kashyap, "Models for motion-based video indexing and retrieval," IEEE Trans. Image Processing, vol. 9, pp. 88–101, Jan. 2000.

[2] N. Moenne-Loccoz, E. Bruno, and S. Marchand-Maillet, "Local feature trajectories for efficient event-based indexing of video sequences," In Proc. of International Conference on Image and Video Retrieval (CIVR), pages 82-91, Tempe, AZ, July 2006.

[3] R. V. Babu and K. R. Ramakrishnan, "Compressed domain motion segmentation for video object extraction," in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 4, pp. 3788–3791, 2002.

[4] R. V. Babu, K. R. Ramakrishnan, and S. H. Srinivasan, "Video object segmentation: a compressed domain approach," IEEE Trans. Circuits. Syst. Video Technol., vol. 14, no. 4, pp. 462-474, 2004.

[5] C.C. Hsieh; W.R. Lai; A. Chiang, "A Real Time Spatial/Temporal/Motion Integrated Surveillance System in Compressed Domain", Intelligent Systems Design and Applications (ISDA), Vol 3, pp:658-665, 2008

[6] G. Piriou, P. Bouthemy, J. F. Yao; "Recognition of Dynamic Video Contents With Global Probabilistic Models of Visual Motion", IEEE Trans. on Image Processing, Vol 15, Issue 11, pp:3417-3430,2006

[7] T. Yu and Y. Zhang, "Retrieval of video clips using global motion information," Electron. Lett., vol. 37, no. 14, pp. 893–895, July 2001.

[8] V. Mezaris, I. Kompatsiaris, N. V. Boulgouris, and M. G. Strintzis, "Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval," IEEE Trans. Circuits. Syst. Video Technol., vol. 14, no. 5, pp. 606–621, May 2004.

[9] M. Ritch and N. Canagarajah, "Motion-based video object tracking in the compressed domain", in Proc. IEEE Int. Conf. on Image Processing(ICIP), vol. 6, pp:301-304, Sept. 2007.

[10] R. Ewerth, M. Schwalb, P. Tessmann, and B. Freisleben, "Segmenting moving objects in MPEG videos in the presence of camera motion", In Proc. of International Conference on Image Analysis and Processing (ICIAP), pp:819-824, Sept. 2007.

[11] Xingquan Zhu, Xiangyang Xue, Hangzai Luo, and Lide Wu, "A Qualitive Camera Motion Classification Based on Motion Vector," Journal of Computer Research And Development (in Chinese), Vol. 38, No. 1, Jan. 2001.

[12] B. L. Yeo and B. Liu, "On the extraction of DC sequences from MPEG. compressed video," in Proc. Int. Conf. Image Processing, vol. II, pp. 260–263, 1995

[13] M. Sonka, V. Hlavac, and R. Boyle, "Image Processing, Analysis, and Machine Vision (2nd Edition)," pp.259-262, PWS Pub., New York, 1999

[14] TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid/